Clustering Longitudinal Unbalanced Data: An Application to the Early Childhood Growth Pattern

Md Jobayer Hossain[1,2]

[1]Biostatistics Core, Department of Biomedical Research, Nemours/Alfred I. duPont Hospital for Children, 1600 Rockland Road, Wilmington, DE 19803

[2]Adjunct Associate Professor, Department of Applied Economics and Statistics, University of Delaware, 213 Townsend Hall, Newark, DE 19716

Abstract

Investigating the presence of distinct trajectories, forming groups of individuals with similar trajectories, and identifying individual and group-level factors contributive to distinct trends are areas of growing interest in longitudinal data analysis. Methods based on dissimilarity in shapes across trajectories are mainly used for clustering unbalanced longitudinal data. Model-based approximation of curves improves precision for sparse and irregularly spaced data. This paper used the empirical best linear unbiased prediction (BLUP) of random coefficients in the piecewise linear mixed effects model for approximating curves, and then used heuristic as well as model-based algorithms for clustering BLUP and their functionally transformed scores. To select an optimum cluster solution, resulting clusters were evaluated in a model-based fashion by adding one cluster and its interaction with time variables at a time to the above piecewise linear mixed effects model. When applied to a dataset of 3365 children aged 1–60 months with a cubic polynomial population growth curve, this clustering technique identified a grouping of six distinct growth trajectories as the best solution.

Keywords: Clustering; Unbalanced; Longitudinal Data; Subject-specific; Trajectory; Vertical Level; Shape

1. Background and Rationale

In unbalanced longitudinal studies, each individual consists of a vector of measurements of an outcome variable for a sequence of irregularly spaced time points. Number of occasions of measurements differs greatly across individuals. Repeated measurements of the same individual are correlated, but vectors of measurements between individuals are independent. The main objectives in the analysis of longitudinal data are to study the individual- and population-level changes in mean response trajectories over time and their relationship with influential covariates. The over-time response trajectories can vary substantially across individuals, and the growing interests in the analysis of these data are investigating the existence of distinct patterns in the trajectories of an outcome of interest; forming groups of homogeneous trajectories; and exploring individual- as well as group-level factors contributive to the varied patterns of trajectories. Conventional cluster algorithms for ordinary multivariable data are not directly applicable for grouping unbalanced longitudinal data without some kinds of adaption or adjustment.

A typical approach to clustering these data is to express the repeated measures on the same individual in simple parametric and nonparametric curves and then to classify these curves into homogeneous groups using a suitable similarity measure of shapes and vertical levels. Usually, some kind of spline basis is used to fit each curve and then traditional heuristic clustering algorithms are applied to the basis coefficients or model-based approaches for clustering are used[1-4]. One difficulty of this approach is

approximating the curve with minimum bias and variance. When data are dense, approximation of each curve based on the data from the corresponding individual could be acceptable. However, for sparse data, this can produce large bias and variability. To overcome the difficulty involving fitting curves on sparse data, James et al used a random effects model with a cubic spline basis allowing projection of curves to borrow strength from the data across individuals and used a mixer model for clustering[4]. Alternatively, Chiou et al used the correlation between random functions for measuring similarity between curves[5]; however, this method ignores the vertical level, which is important in clinical studies. Nagin et al developed a group-based mixture model approach for clustering longitudinal trajectories, but this method is not flexible enough to handle sparse and irregularly spaced data[6-7].

This article aimed to use the empirical best linear unbiased prediction (eBLUP) of random coefficients in the piecewise mixed effects model to approximate the heterogeneity in the vertical levels and shapes across all individuals under study and then to apply conventional heuristic as well as model-based approaches of cluster algorithms to classify individuals with similar patterns of trajectories in the same group and with varied patterns in distinct groups. The rationale of using the linear mixed effects model for approximating curves is pragmatic as this model is a well known, effective statistical technique for modeling mean response trajectories in unbalanced longitudinal data as a combination of population and subject-specific effects. The model expresses the time dependence of repeated measures as a function of time and thereby handles the unbalanced longitudinal data with relatively few parameters irrespective of the number of timing of measurements[8]. The piecewise linear mixed effects model simplifies the shape of a trajectory of complex polynomials by expressing in slopes of a sequence of line segments. In addition to the precision in modeling, the use of this model reduces computational burden substantially as it is available in most statistical software packages. As for pre-processing of cluster inputs, this method allows using scores of functional transformations of eBLUP for clustering. These scores are modifications of BLUP prior to clustering and usually represent a natural structure of distinct trajectories in smaller dimensions, thereby aiding in better clustering. The method described in this paper also provides a model-based evaluation to select a cluster solution that provides an optimal representation of the structure of trajectories in the longitudinal unbalanced data.

The rest of the article is organized as follows: section 2 presents a brief review of the piecewise mixed effects model and the best linear unbiased prediction of random coefficients; section three discusses clustering analysis of eBLUP and the selection of an optimum solution; section 4 appliesthis novel method of clustering to the early childhood growth trajectories; and section 5 provides discussion and concluding remarks.

2. Review of Piecewise Linear Mixed Effects Models and the Best Linear Unbiased Prediction (BLUP)

2.1. Linear Mixed Effects Model

The general form of the mixed effects model can be written as,

(1): $Y_i = X_i\beta + Z_i b_i + e_i$,

where Y_i is the $(n_i \times 1)$ vector of the repeated responses for the *i*th individual, β is a $(p \times 1)$ vector of fixed effects associated with $(n_i \times p)$ design matrix of covariates X_i, b_i

is a $(q \times 1)$ vector of random effects associated with $(n_i \times q)$ design matrix $Z_i \subset X_i$ with $q \leq p$. Z_i comprises the covariates of X_i, for those corresponding components of β can vary randomly from one individual to another. The population characteristics, β, that are shared by all individuals under study, link X_i to Y_i. The subject-specific effects, b_i, illustrate how regression parameters for ith individual corresponding to Z_i deviate from that of β. Thereby, b_i, accounts for the heterogeneity across all individuals under study. The effects, b_i, are assumed to be distributed as multivariate normal with mean 0 and covariance matrix G, i.e. $b_i \sim MN(0, G)$. Similarly, the $(n_i \times 1)$ vector of errors, e_i, is assumed to be distributed as multivariate normal with mean 0 and covariance matrix R_i, i.e. $e_i \sim MN(0, R_i)$. Generally, R_i is assumed to be a diagonal matrix, ${}^2I_{n_i}$. The vectors b_i and e_i are assumed to be independent.

Henderson[9] derived the best linear unbiased estimates (BLUE) of β and the best linear unbiased prediction (BLUP) of b_i. For k indivduals, the BLUE of β is,

$$\hat{\beta} = \left\{ \sum_{i=1}^{k} (X_i' V_i^{-1} X_i) \right\}^{-1} \sum_{i=1}^{k} (X_i' V_i^{-1} Y_i) \text{ with } Var(\hat{\beta}) = \left\{ \sum_{i=1}^{k} (X_i' V_i^{-1} X_i) \right\}^{-1}.$$

As $$\begin{pmatrix} Y_i \\ b_i \end{pmatrix} \sim MN\left(\begin{pmatrix} X_i \beta \\ 0 \end{pmatrix}, \begin{pmatrix} V_i = Z_i G Z_i' + R_i & GZ_i \\ Z_i'' G & G \end{pmatrix} \right),$$

by definition of the conditional mean of two multivariate normal variables, the BLUP of b_i for given Y_i can be given as

$$\hat{b}_i = E(b_i | Y_i) = G Z_i V_i^{-1} (Y_i - X_i \hat{\beta})$$

with the prediction error,

$$Cov(\hat{b}_i - b) = G - G Z_i^{-1} V_i^{-1} Z_i G + G Z_i^{-1} V_i^{-1} X_i \left\{ \sum_{i=1}^{k} (X_i' V_i^{-1} X_i) \right\}^{-1} X_i V_i^{-1} Z_i G.$$

In BLUP, $\hat{b}_i$, V_i and G are unknown and can be replaced by their maximum likelihood (ML) or restricted ML (REML) estimates. This later expression of $\hat{b}_i$ is known as the empirical BLUP (eBLUP). The same prediction can be derived from Henderson's mixed model equation (MME). Once we know $\hat{\beta}$ and $\hat{b}_i$, the trajectory, $\hat{Y}_i$ for given subject specific characteristics of the ith subject is obvious. Specifically, $E(Y_i | b_i) = \hat{\beta} X_i + \hat{b}_i Z_i$. This is known as the subject-specific trajectory, while the population trajectory is $E(Y) = \hat{\beta} X$.

2.2 Piecewise Linear Mixed Effects Model

The form of the time dependence of repeated measures of an individual can be linear or non-linear. The simplest curve is the straight line. Only an intercept and a slope are sufficient to describe this curve, and the slope has simple interpretation in terms of the

constant rate of change in the mean response. In many applications, longitudinal data change in irregular rates over time. Non-linear trends of this type may not be well approximated by polynomials of any order. One approach to represent polynomial curves of this type is to have a sequence of connected line segments that produces a piecewise linear trend of repeated measures with different slopes in different segments but joined together at fixed times. Thus, the piecewise linear mixed effects model provides us with a simple representation of population- and subject-level mean temporal changes of any order of polynomial in terms of an intercept and slopes of a sequence of line segments. The mixed effects model in equation (1) can be transformed to the linear piecewise form by multiplying design matrices X and Z by a matrix S of linear spline basis so that $SX = X^*$ and $SZ = Z^*$. Then, the piecewise form of the model in the equation (1) becomes

$$(2): \quad Y_i = X_i^* \beta + Z_i^* b_i + e_i.$$

Let us consider a model of pth polynomial of time. Considering only the time variables in the model, design matrix X can be transformed to X^* as follows,

$$X_i = \begin{pmatrix} 1 & t_{i1} & \cdots & t_{i1}^p \\ 1 & t_{i2} & \ldots & t_{i2}^p \\ \vdots & \vdots & \vdots & \vdots \\ 1 & t_{ini} & \cdots & t_{ini}^p \end{pmatrix}, \; X_i^* = \begin{pmatrix} 1 & t_{i1} & t_{i1} - t_1^* & \cdots & t_{i1} - t_m^* \\ 1 & t_{i2} & t_{i2} - t_1^* & \ldots & t_{i2} - t_m^* \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & t_{ini} & t_{ini} - t_1^* & \cdots & t_{ini} - t_m^* \end{pmatrix}$$

Where, $(t_{ij} - t_m^*)_+ = \begin{cases} t_{ij} - t_m^* & \text{for } t_{ij} - t_m^* > 0 \\ 0 & \text{Otherwise;} \end{cases}$

Where, t_{ij} is the time of the jth measurement for the ith subject; the number of knots, $m = p - 1$, is one less than the number of segments. Usually, a pth polynomial of time can be represented by p linear segments of time. Similarly, the design matrix Z can be transformed to Z^*. The estimation of β and the prediction of b_i are the same as before.

3. Cluster Analysis of BLUP and Selection of the Optimum Solution

The analysis of the piecewise mixed effects model in equation (2) provides the predicted values of random coefficients, $b_{il}, i = 1,2, \ldots\ k, l = 1,2, .. m, m+1, m+2$. The BLUP $\hat{b}_{il}, i = 1,2, \ldots\ k, l = 1,2, .. m, m+1, m+2,$ can be arranged in $(m+2)$ time-ordered variables with k rows for each variable. Each of the k components of the first time-ordered variable, $\hat{b}_{i1},$ measures the variability in the vertical level for each individual at the start of the study, and rows of the second to (m+2)*th* variables measure the variability in shapes of the curves. As no algorithms uniformly work well on all datasets, a large number of clustering approaches—ranging from heuristic approaches, such as hierarchical algorithms[10] and *k*-means-like partitioning algorithms[11], to formal model based approaches with classification and mixture likelihood[12]—can be applied for clustering BLUP. Some of the conventional clustering methods may fail to recognize time direction of BLUP data during clustering[13]. Thus, in addition to BLUP data, scores from functional transformations, such as principal component, factor and canonical analyses, can be used[14]. In addition, a large number of parametric and non-parametric variable standardization techniques can be used. After a series of cluster solutions are produced using different cluster algorithms and inputs, the next step is to select the

optimum cluster solution. There are many methods for internal and external validations of cluster analysis, including cross validations, bootstrapping, mixer model, and non-parametric density estimation[15]. However, a meaningful clustering would be able to extract the natural structure of the data in line with the purpose of the data analysis. Considering this fact, it is reasonable to use a model that extends equation (2) by including the group variable under evaluation and its interaction with time variable. This process of the evaluation will be repeated for all cluster solutions. The group variable that fits the data best can be accepted as the optimum solution.

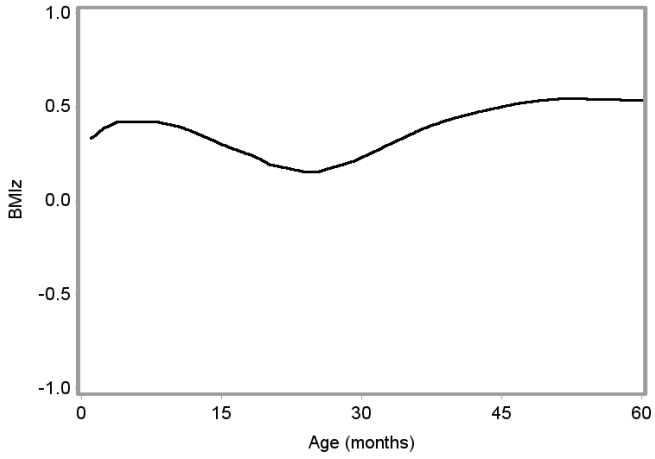
4. Application to Early Childhood Growth Pattern

4.1 Data Description

The dataset consists of the standardized scores of weight-for-length (ages < 2 years) and body mass index (BMI) (ages between 2 and 5 years) from 3365 children collected on their clinic visits during the first 5 years of life. Standardized scores of these two variables are termed as BMIz and used as the early childhood growth indicator in this article. The data were retrieved from the Nemours electronic health records after approval of the Nemours Institutional Review Board. Children who had the first visit at any Nemours clinics by the first month of the birth and thereafter had at least one visit each year for the next 5 years of life were included in the study. Children with cancer and cystic fibrosis were excluded from the study because of the potential abnormal growth pattern. Children visited clinics on their own health care purposes. Thus, each child had a unique sequence of clinic visits and the resultant dataset was extremely unbalanced with sparse and irregularly spaced measurements within and across the children. Analysis included the measurements of BMIz between ages 1 and 60 months. Measurements at ages 1 and 60 months were interpolated when needed. A total of 51711 clinic visits were made by 3365 children. The median (IQR) number of visits was 14 (12–17) with a range of 6–76.

4.2 Visual Inspection and Model Consideration

Figure 1: Loess Smooth Trend in Childhood BMIz (1-60 Months)



The LOESS smooth curve in Figure 1 depicts the salient feature of the mean change in BMIz as a function of time. It reveals a form of cubic polynomial in the change in BMIz over time during the first 5 years of life that can be approximated by three linear trends with varying slopes at three different segments of ages. That is, a model is needed that

describes each child's growth curve with an intercept and three slopes at three different segments. An examination of a piecewise linear mixed effects model with random coefficients exhibited the best fit of this dataset with knots at 8 and 21 months. Thus, the three segments of ages that contain approximate linear trends in the temporal change in BMIz are 1–8 months, 8–21 months, and 21–60 months. The growth trajectories of the ith child can be represented by the following piecewise linear mixed effects model:

(3): $E(Y_{ij}|b_i) = \beta_0 + \beta_1 t_{ij} + \beta_2(t_{ij} - t_1^*) + \beta_3(t_{ij} - t_2^*) + b_{i0} + b_1 t_{ij} + b_{i2}(t_{ij} - t_1^*) + b_{i3}(t_{ij} - t_2^*)$

Where, $(t_{ij} - t_m^*)_+ = \begin{cases} t_{ij} - t_m^* & \text{for } t_{ij} - t_m^* > 0 \\ 0 & \text{Otherwise} \end{cases}; m = 1, 2; i = 1,2....3365; j = 1,2,....k_i.$

In the model above, Y_{ij} denotes the BMIz of the ith child at the jth measurement and $t_{ij} = Age_{ij} - 1$ denotes the time of the corresponding measurement, $t_{ij} \in [0, 59]$. Two knots are at $t_1^* = 7$ and $t_2^* = 20$. $\beta = [\beta_0, \beta_1, \beta_2, \beta_3]$ are the regression coefficients of the population characteristics and $b_i = [b_{i0}, b_{i1}, b_{i2}, b_{i3}], i = 1,2, \ldots.3365,$ are the ith child's deviation from the corresponding component of β. β_0 and $\beta_1$represent the population intercept or the mean BMIz at the age of 1 month and the rate of change in BMIz (slope) during the ages between 1 and 8 months (first segment), respectively. β_2 and β_3 are the differences between slopes in the first and second segments and the second and third segments, respectively. The individual level regression coefficients,

b_{i0}, b_{i1}, b_{i2} and b_{i3} $i = 1,2, \ldots.3365,$ account for the heterogeneity in levels and shapes of the trajectories among 3365 children. Table 1 presents the distribution of the predicted four random coefficients. There are substantial variability among children in predicted values of b_{i0}, b_{i1}, b_{i2} and b_{i3}, $i = 1,2, \ldots.3365, p < 0.0001$ for all four variables of coefficients.

Table 1: Distribution of the BLUP of Random Coefficients

Coefficients	Mean (SE)	Median (IQR)	Min, Max	Variance (SE)
Intercept	0.00(0.01)	-0.48 (0.03, 0.52)	-3.17, 2.37	0.69 (0.02)
Slope in Segment 1	0.00(0.01)	-0.81 (-0.03, 0.78)	-4.76, 6.36	2.12 (0.07)
Slope in Segment 2	0.00(0.01)	-0.93 (-0.01, 0.91)	-6.77, 6.82	-2.36 (0.11)
Slope in Segment 3	0.00(0.01)	-0.38 (0.03, 0.42)	-3.13, 2.3	-0.69 (0.02)

4.3. Clustering Four Empirical BLUP

As discussed in section 3, a large number of clustering algorithms are available in most of the popular statistical software packages, along with various options of distance or similarity measures. A number of cluster solutions of different sizes are generated using cluster algorithms, such as hierarchical, k-means, non-parametric density, finite Gaussian mixture modeling, and combinations of methods in two or more steps that are available in SAS, R, and SPSS. Predicted coefficients as well as their principal component (PC) and

factor scores are used as input for clustering. The first three PCs explained 99.7% of the variation in 4 variables of random coefficients and are selected for clustering. Similarly, the scores of the three factors are used for clustering. To select an optimum cluster solution that classifies children with similar trajectories in the same group and has a number of distinct groups that adequately represent the natural structure of the data, each cluster solution and its interaction with time variables were added in the model of equation (1). Thus, the cluster solution in the following model that fits the longitudinal BMIz data best can be accepted as the optimum solution.

(4): $E(Y_{ij}|b_i) = \beta_0 + \beta_1 t_{ij} + \beta_2(t_{ij} - t_1^*) + \beta_3(t_{ij} - t_2^*) + \beta_4 ClusterGroup + \beta_5 ClusterGroup * t_{ij} + \beta_6 ClusterGroup * (t_{ij} - t_1^*) + \beta_7 ClusterGroup * (t_{ij} - t_2^*) + b_{i0} + b_1 t_{ij} + b_{i2}(t_{ij} - t_1^*) + b_{i3}(t_{ij} - t_2^*)$ …….(1)

Where, $(t_{ij} - t_m^*)_+ = \begin{cases} t_{ij} - t_m^* & \text{for } t_{ij} - t_m^* > 0 \\ 0 & \text{Otherwise} \end{cases}; m = 1, 2; i = 1,2....3365; j = 1,2,....n_i.$

Model fit criteria BIC and AIC are used to select with optimum cluster groupings. Table 2 presents the values of BIC and AIC for evaluating cluster solutions of sizes 4–6 using eBLUP and their PC and factor scores as cluster inputs.

Table 2: Model Fit Statistics for Cluster Evaluation

Cluster Input	Number of Clusters	BIC	AIC
Factor Scores	6	94495.8	94538.7
PC scores	6	94649.9	94692.8
PC scores	4	94851.7	94894.5
PC scores	5	94869	94911.8
BLUP	6	95077.8	95120.6
BLUP	4	95098.6	95141.5
BLUP	5	95325.2	95368.1
Factor Scores	4	95862.6	95905.4
Factor Scores	5	95932	95974.8

The solution of cluster size 6 using factor scores as the input generated by a two-step hierarchical algorithm yielded the minimum BIC and AIC, indicating superior results compared to others.

Figure 2: Distinct Early Childhood Growth Trajectories

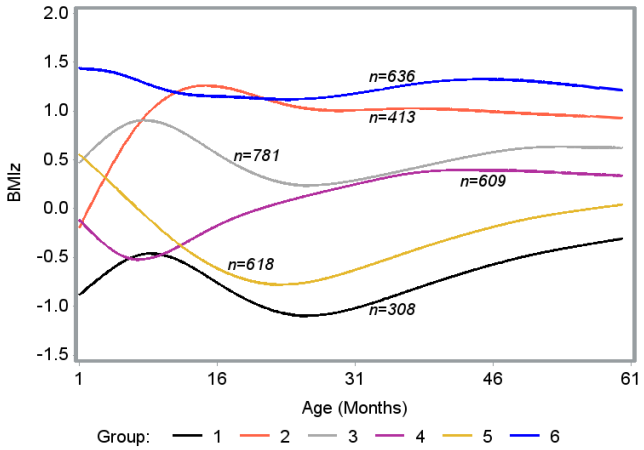


Figure 2 displays the LOESS smooth trajectories of early childhood BMIz for six distinct groups. The figure also presents the number of children in the corresponding groups.

5. Discussion and Conclusions

Clustering of longitudinal data collected in irregularly spaced time points for each individual involves several steps of analysis. Application of efficient methods in each phases of analysis is essential for effective and useful groupings. The first step of the analysis is to approximate the longitudinal trajectory or time dependence of repeated measures of each individual with sufficient accuracy. The second step is the use of a suitable clustering algorithm for the detection of similarity or dissimilarity in trajectories among all individuals under study. The last step is the evaluation of the cluster results in order to select an optimum grouping of trajectories. This article uses a piecewise linear mixed effects model to approximate the variability in the vertical shift and the shape of the trajectories. As described in sections 1 and 2, the model uses population and subject-specific coefficients for approximating the trajectory of an individual. Intercepts approximate the vertical level at the beginning of the curve, while the slopes of piecewise splines approximate the shapes of the trajectory. The model offers the unbiased estimation as well as unbiased prediction of the population and subject-specific regression coefficients with the least possible variance and prediction error, respectively. The advantages of the piecewise spline are discussed in section 2. Compared to individual data-based curve projection, this model allows us to borrow strength across individuals when approximating curves, resulting in superior projection for sparse and irregularly spaced repeated measures. The ability to cluster either predicted random coefficients or scores of their functional transformations gives the flexibility to use any suitable algorithm, including heuristic methods (e.g., hierarchical and *k*-means), model-based methods (i.e., classification likelihood and mixture likelihood approaches), and non-parametric density estimation based methods. Functional transformations of coefficients facilitate extraction of natural structures in the dataset. There is no computational burden to implement the proposed clustering method because linear mixed effects models and conventional clustering algorithms are available in most statistical software packages.

Earlier methods tended to project each curve on the data from the corresponding individual and were highly likely to result in biased and misleading projections. To overcome this drawback, James et al used a mixture likelihood based approach with a

random effects model to fit cubic splines; however, because this method involves heavy computational requirements, it has not been used frequently. Nagin et al used mixture likelihood for group-based probability clustering. The capacity is limited in handling irregularly spaced repeated measures for each individual. Their software is capable of using, at most, a cubic polynomial curve.

The last phase of the analysis involves the evaluation of cluster results and determination of the optimal number of cluster groups. Existing conventional methods are mostly applicable to a dataset with a single row for each individual. Model-based methods with the capability of recognizing the natural structure of longitudinal curves could be suitable. The piecewise mixed effects model in the equation (4) used in this article for cluster evaluation adequately recognizes the longitudinal curves. A group variable with maximum within-group homogeneity and between-group heterogeneity in curves is likely to offer the optimal fit of the proposed model.

The application of the proposed clustering method on the early childhood growth data effectively identified 6 distinct groups of BMIz trajectories with varying vertical levels and shapes of curves. The method also identified latent factor scores as the better input for clustering this particular dataset. In conclusion, this method possesses the desirable capacity to cluster longitudinal data with sparse and irregularly spaced repeated measures.

Acknowledgments

I express special appreciation to Dustin Samples for his meticulous review and editing services. The author received partial salary support from the National Institutes of Health (NIH) COBRE Grant 8P20GM103464-9 (PI: Shaffer) and an Institutional Development Award (IDeA) from the National Institute of General Medical Sciences of the NIH under Grant no. U54-GM104941 (PI: Binder-Macleod).

References

1. Abraham, C., Cornillon, P.A., Matzner-Løber, E.R.I.C., Molinari, N., 2003. Unsupervised curve clustering using B-splines. Scandinavian journal of statistics, 30(3), pp.581-595.

2. Garcia-Escudero, L.A.,Gordaliza, A., 2005. A proposal for robust curve clustering. Journal of classification, 22(2), pp.185-201.

3. Serban, N.,Wasserman, L., 2012. Cats. Journal of the American Statistical Association.

4. James, G.M., Sugar, C.A., 2003. Clustering for sparsely sampled functional data. Journal of the American Statistical Association, 98(462), pp.397-408.

5. Chiou, J.M., Li, P.L., 2012. Correlation-based functional clustering via subspace projection. Journal of the American Statistical Association.

6. Nagin, Daniel S. 1999. "Analyzing Developmental Trajectories: A Semi-Parametric, Group-Based Approach." Psychological Methods 4:139-77.

7. Jones, B.L., Nagin, D.S., Roeder, K., 2001. A SAS procedure based on mixture models for estimating developmental trajectories. Sociological methods & research, 29(3), pp.374-393.

8. Fitzmaurice M.F., GKaufman, Laird N.M., Ware J.H. 2004. Applied Longitudinal Analysis. New York: Wiley-Interscience.

9. Henderson, C.R., 1985. Best linear unbiased prediction of nonadditive genetic merits in noninbred populations. Journal of animal science, 60(1), pp.111-117.

10. Kaufman, L. and Rousseeuw, P. (1990). Finding Groups in Data: An Introduction to Cluster Analysis. New York: Wiley.

11. Hartigan, J.A. and Wong, M.A., 1979. Algorithm AS 136: A k-means clustering algorithm. Journal of the Royal Statistical Society. Series C (Applied Statistics), 28(1), pp.100-108.

12. Banfield, J.D. and Raftery, A.E., 1993. Model-based Gaussian and non-Gaussian clustering. Biometrics, pp.803-821.

13. Luan, Y. and Li, H., 2003. Clustering of time-course gene expression data using a mixed-effects model with B-splines. Bioinformatics, 19(4), pp.474-482.

14. Combes, C. and Azema, J., 2013. Clustering using principal component analysis applied to autonomy–disability of elderly people. Decision Support Systems, 55(2), pp.578-586.

15. Fraley, Chris and Adrian E. Raftery. 1998. "How Many Clusters? Which Clustering Method? Answers Via Model-Based Cluster Analysis." Computer Journal 41:578-88.